面向主题检索的科技政策扩散识别方法

曾立英[1]，　许乾坤[2]，　张丽颖[3]，　刘　耀[2]

(1. 中央民族大学 国际教育学院　北京 100081；2. 中国科学技术信息研究所　北京 100038；
3. 河北省工业和信息化厅　河北 石家庄 050017)

摘要：为满足用户对某一主题下科技政策扩散关系挖掘的分析需求，构建了科技政策扩散识别模型。通过从科技政策文本中提取组织结构相似性特征、语义结构相似性特征、关键词承继性特征及基于 Doc2vec 的文本相似性特征，实现多个特征的一体化处理；提出了基于识别模型评分的科技政策文本距离计算方法，根据政策间的文本距离与扩散概率的关系，寻找使扩散关系成立的最大文本距离，并将扩散经验值融入识别模型中，实现检索过程中科技政策扩散对和扩散集的自动计算和输出。实验结果表明，所构建的科技政策扩散识别框架能有效地提取出扩散集合。

关键词：科技政策；主题检索；政策扩散；文本挖掘；文本分析；扩散特征

中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1671-6841(2022)05-0082-08

DOI：10.13705/j.issn.1671-6841.2022060

Identification Method for Subject Retrieval of Science and Technology Policy Diffusion

ZENG Liying[1], XU Qiankun[2], ZHANG Liying[3], LIU Yao[2]

(1. *College of International Education, Minzu University of China, Beijing* 100081, *China*;
2. *Institute of Scientific and Technical Information of China, Beijing* 100038, *China*;
3. *Industry and Information Technology Department of Hebei Province, Shijiazhuang* 050017, *China*)

Abstract: In order to meet the users′ analysis requirements for mining the diffusion relationship of science and technology policies of a certain theme, a identification model of science and technology policy diffusion was constructed. By extracting organizational structure similarity features, semantic structure similarity features, keyword inheritance features, and text similarity features based on Doc2vec from science and technology policy texts; the integration of multiple features could be realized. Then, a method to calculate the text distance of science and technology policy based on the recognition model score was proposed. According to the relationship between the text distance and diffusion probability between policies, the maximum text distance that made the diffusion relation hold was found; and the diffusion experience value was integrated into the recognition model to realize the automatic calculation and output of the diffusion pair and diffusion set of science and technology policy in the retrieval process. The experimental results showed that the established framework of science and technology policy diffusion identification could effectively extract diffusion sets.

Key words: science and technology policy; subject retrieval; policy diffusion; text mining; text analysis; diffusion characteristics

收稿日期：2022-03-12
基金项目：国家社会科学基金项目(21BTQ011)；国家重点研发计划项目(2018YFB143502)。
第一作者：曾立英(1970—)，女，教授，主要从事计算语言学和汉语语言学研究，E-mail：lizzengliying@qq.com。
通信作者：刘耀(1972—)，男，研究员，主要从事自然语言处理和知识工程研究，E-mail：liuy@istic.ac.cn。

0　引言

政策扩散最早出现于 1969 年[1],这一时期的研究内容主要包括政策扩散的定义、扩散现象的研究方法及扩散的原因。在大数据背景下,传统的政策文本分析方法已不能满足政策研究的需要。作为一种新的政策文本解读方式[2],政策文本计算能将自然语言处理领域的文本分析和计算与传统的社会科学研究结合在一起。现有的政策扩散研究中,针对同一主题的政策扩散研究是当前的热点问题。张剑等[3]利用构建中国公共政策参照网络,抽取关键主题进行时序分析,提出新的公共政策扩散的文献量化研究维度与方法。裴雷等[4]通过信息化政策地域扩散和历时扩散两个维度的测度,分别对信息化政策扩散中的主题承继与主题创新、主题跃迁与主题衰退、政策扩散涟漪效应与漏斗效应的显著性进行了检测验证。武学振[5]提出了政策创新扩散指数,通过计算省级政府政策发布时滞与同主题政策发布最大时滞的比值,对各省信息政策扩散能力的强弱进行判断。

Grimmer 等[6]认为,政策文本分析的起点是获取政策文本,构建政策文本库,并提出了有效收集新文本的方法。李江等[7]提出了文献量化分析方法向政策文本的迁移,许多针对学术文献的分析方法如时间序列分析、共词分析及网络分析等,都可以用于政策文本的量化分析。此外,Nowlin[8]基于 LDA 方法提出了政策问题定义模型。但是,由于政策文本内容结构复杂,基于单特征的方法只能对部分政策扩散关系进行判定。例如,王小杰[9]考虑了政策间的参照关系特征,并将文本中出现了参照关系的政策视为扩散政策。李庆[10]通过人工编码的方式将政策内容进行主轴编码和开放编码,并通过计算相同编码的比例来判断政策扩散关系。

对于如何获取最贴近该主题的扩散集合,如何区分不同子主题的扩散集合和扩散现象等基本问题,研究者们却很少涉及。基于上述分析,本文从文本结构和语义特征提出了多特征的科技政策扩散关系判定方法,并基于机器学习中的排序算法,构建了科技政策扩散识别模型,用来发掘与主题相关性较高的扩散源及对应的扩散政策,最终实现检索过程中科技政策扩散对和扩散集的自动计算和输出。

1　扩散识别模型的构建

面向主题检索任务,构建了科技政策扩散识别框架,发掘与主题相关性较高的扩散源及对应的扩散政策,并建立面向不同子主题的扩散对和扩散簇。通过借鉴排序学习的思想,构建融合多样化特征的识别模型,使输出的扩散集合能对不同子主题下的科技政策进行覆盖。同时,基于扩散概率与文本距离的相关性,将扩散关系计算转换为文本距离计算,并以此来优化识别模型。科技政策扩散识别模型的研究框架如图 1 所示。

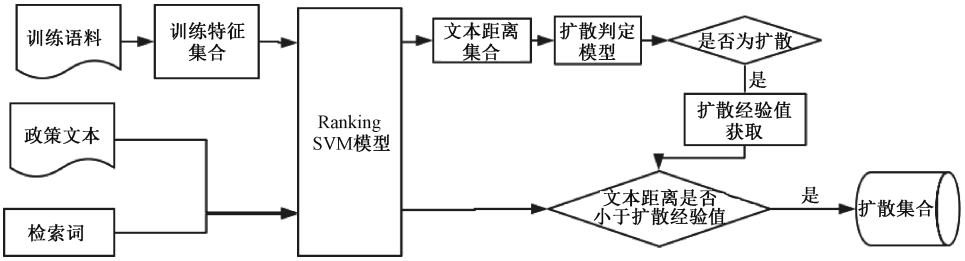

图 1　科技政策扩散识别模型的研究框架

Figure 1　Research framework of science and technology policy diffusion identification model

政策扩散特征分为政策结构特征和政策内容特征,其中政策结构特征包括组织结构相似性特征与语义结构相似性特征,政策内容特征包括关键词承继性特征与文本相似性特征。特征确定后,需要从文本中提取上述特征并进行表示。文本特征提取主要有基于规则和基于统计两种方法,本文主要采用基于统计的方法,将特征表示为实数向量,以便将其应用于分类模型。

1.1　组织结构相似性特征

政策组织结构是定义政策问题及解释因果的关键。如果两篇政策存在扩散关系,那么这两者之间很可能存在组织结构的联系,可以将两篇政策的结构联系程度作为扩散的结构特征。同时,通过对政策文本进行分析,发现当扩散政策继承了扩散源对问题的定义和解决方法时,扩散源中的组织结构特征词应在扩散政策同等层级中出现。

扩散源与扩散政策结构对照如图 2 所示。其中

左侧为扩散源,右侧为扩散政策,二者除政策标题基本一致外,同一层级的特征词,例如“科技成果转化”“融资渠道”“实体经济转型”“人才流动”等也基本相同。

本文所提出的科技政策语义结构就是利用政策领域词表对科技政策原有组织结构进行重组,使其能够在同一标准下与其他政策文本进行比较。组织结构提取流程如图3所示。

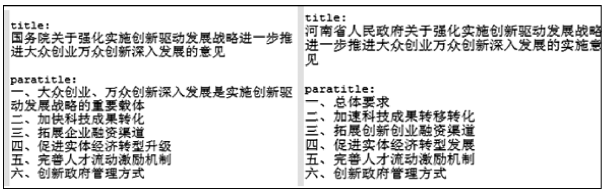

图2 扩散源与扩散政策结构对照

Figure 2 Comparison of diffusion source and diffusion policy structure

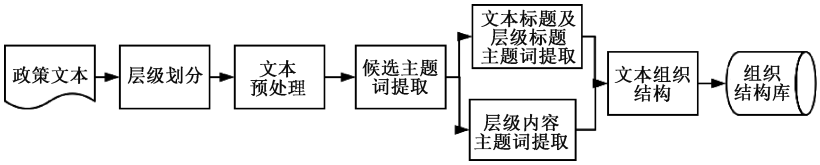

图3 组织结构提取流程

Figure 3 Organization structure extraction process

为了建立两篇政策的组织结构联系,提出了组织结构相似度计算方法。科技政策一般会按照序数符号标明层次,如第一层用“一、”,第二层用“(一)”,第三层用“1.”,第四层用“(1)”。各层级的首句定义为该层级的标题,将政策标题与第一层标题的内容定义为该篇政策文本的摘要。将两篇政策中较早的政策设为候选扩散源,其组织结构表示为 $S=(V,L)$,其中 $V=\{v_1,v_2,\cdots,v_i\}$,代表扩散源组织结构中的特征词集合。

两篇政策的组织结构相似度为

$$WO(d^A,d^B)=\sum_{i=1}^{n}\left(\rho_i\times\frac{N_{i(A\cap B)}}{N_{iA}}\right),\qquad(1)$$

其中:N_{iA} 为第 i 层政策 A 中组织结构特征词数量;$N_{i(A\cap B)}$ 为第 i 层政策 A 与文档 B 重合的特征词数量;n 表示文章划分的层级数,$n\leqslant 5$,第0级为文章标题;ρ_i 为层级 i 的权重。层级越高,权重越大,因此将权重设为

$$\rho_i=(n-i)/\sum_{i=1}^{n}i。\qquad(2)$$

1.2 语义结构相似性特征

语义结构可以体现扩散源与扩散政策在整个科技政策体系中的联系。由于每篇政策的语义结构都是一个树结构,在比较两篇政策语义结构相似性时,可以通过计算扩散政策语义结构对扩散源语义结构的覆盖程度来计算相似度。具有扩散关系的两个语义结构往往比不具有扩散关系的两个语义结构有更多的公共节点,语义结构比较如图4所示,其中阴影部分为相似节点。

两篇政策的语义结构相似度为

$$WS(d^A,d^B)=\frac{\sum_{i=1}^{n}N_{i(A\cap B)}}{\sum_{i=1}^{n}N_{iA}}。\qquad(3)$$

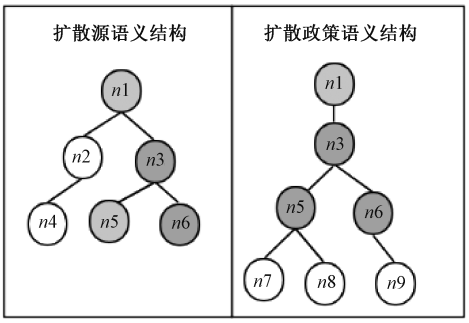

图4 语义结构比较

Figure 4 Comparison of semantic structure

1.3 关键词承继性特征

在政策扩散过程中,相比政策领域通用词汇,政策关键词的扩散概率更高。通过尽可能多地提取政策扩散源的关键词,使其尽量覆盖政策扩散源的内容,并选用合适的特征来表征待分析文本,可以建立其与政策扩散源内容间的广泛联系。本文最终的判定模型为分类模型,需体现特征在不同文本中的差异性。因此,选取 *TF-IDF* 提取每篇文档的关键词特征。

假设两篇政策的关键词集为 $W=\{w_1,w_2,\cdots,w_n\}$,其关键词承继性特征的计算方法为

$$Sim(d^A,d^B)=\sum_{i=1}^{n}\min\{tfidf(w_i^A),tfidf(w_i^B)\},\qquad(4)$$

其中:$tfidf(w_i^A)$ 为关键词集中第 i 个词在文档 A 中的 *TF-IDF* 值。

1.4 基于 Doc2vec 的文本相似性特征

除关键词以外,扩散政策也会采纳政策扩散源的段落内容,采纳内容越多,文本相似度越高,证明越可能发生了政策扩散。因此,需要将提取文本相

似性特征作为判定扩散关系的因素之一。

由于政策文本中不同概念间存在较强的相关性,单纯基于词汇对应的距离计算并不能充分体现文本的相似度。为了能在文本表示中融入更多的语义,以体现出文本间的语义关联,利用 Doc2vec 将文本表示为向量,再通过余弦值计算文本间的相似度。利用 Doc2vec 计算文本相似度的流程如图 5 所示。

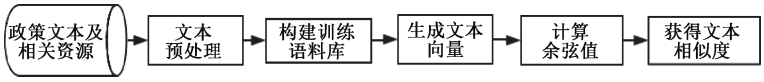

图 5　利用 Doc2vec 计算文本相似度的流程

Figure 5　The process of calculating text similarity with Doc2vec

具体步骤如下。

1) 对科技政策进行预处理,之后构建文本语料库。引入 2021 年各地区科技工作报告以及 2001—2021 年由科学技术信息研究所编写的科技参考,共 1 823 篇,与科技政策共同构建政策文本语料库。

2) 利用语料库训练 Doc2vec,使用 python 的 genism 包对构建好的语料库进行训练。训练过程中参数设置如下:定义特征向量的维度为 100,参与训练的最小词频为 1,滑动窗口的大小为 9,选用 DM 训练算法,迭代次数为 50。

3) 生成每篇政策文本的向量化表示。

4) 利用余弦值计算政策扩散源与其他政策文本的相似度。余弦值在[-1,1]之间,余弦值越高,表明两篇文本越相似。具体计算公式为

$$Sim(v_i,v_j)=\frac{v_i\times v_j}{\|v_i\|\times\|v_j\|}。\tag{5}$$

1.5　识别模型构建

面对一个主题检索任务,文本之间传统的排序方式是按照相关性、出版时间、被引量、下载次数等进行排序。但这些排序往往较为片面,不能有效地覆盖相关主题内容,更不能体现不同政策文本间结构与内容的差异。为了使科技政策扩散识别模型的输出结果既能体现文本与主题的相关性,又能覆盖较多的子主题,体现出检索结果的子主题差异性,选用 Ranking SVM 构建识别模型,拟在对文本组织结构相似性、文本语义结构相似性及文本内容相似性特征提取的基础上,融合多样化特征及主题与政策文本的相关性特征。Ranking SVM 算法将待排序对象组合成文档对,利用 SVM 来解决文档对的排序任务。

给定一组查询 $\{q_1,q_2,\cdots,q_n\}$,对于文档对集合 $\{D_u^i,D_v^i\}$,若文档 D_u^i 的排序高于 D_v^i,则将其对应的值标为 1,反之标为 0,进而获得了文档对的真值标注集合 $y_{u,v}^i$。通过上述过程可以将文档的排序问题转换为分类问题。

定义排序打分函数 $f(x)=\boldsymbol{w}_x^{\mathrm{T}}$,则相应的优化问题可形式化为

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\sum_{u,v;y_{u,v}^i}\xi_{u,v}^i,\tag{6}$$

$$\boldsymbol{w}^{\mathrm{T}}(D_u^i-D_v^i)\geqslant 1-\xi_{u,v}^i,\ \text{if } y_{u,v}^i=1,\tag{7}$$

$$\xi_{u,v}^i\geqslant 0,i=1,2,\cdots,n,\tag{8}$$

其中:$\boldsymbol{w}$ 为参数向量;x 为文档的特征;y 为文档对之间的相对相关性;ξ 为松弛向量。利用 Ranking SVM 算法构建科技政策扩散识别模型的流程如图 6 所示。

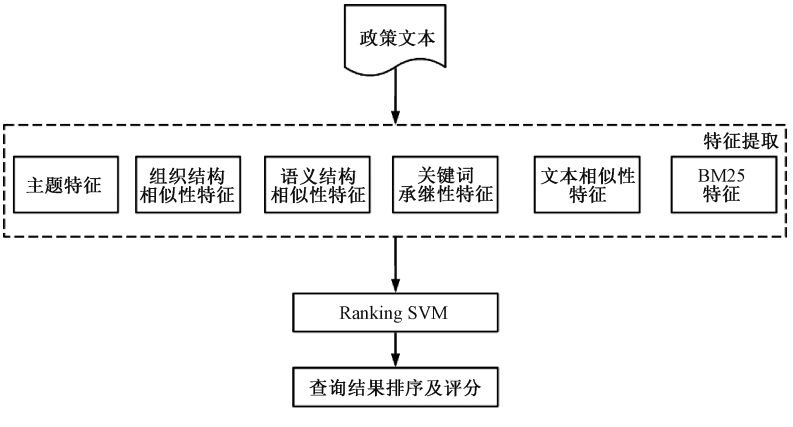

图 6　利用 Ranking SVM 算法构建科技政策扩散识别模型的流程

Figure 6　The process of constructing the identification model of science and technology policy diffusion by using Ranking SVM algorithm

根据科技政策扩散识别模型构建流程,需要获取相关性和多样性特征作为模型的输入。除了前面介绍的特征外,拟使用LDA模型获取主题多样性特征,利用BM25模型获取主题与政策文本相关性特征。

1.5.1 基于LDA模型的主题多样性特征获取 LDA是一个具有文档层、隐藏主题层以及特征词层的三层贝叶斯模型。将语料库中的每篇文献视为隐藏主题的概率分布,同时每个隐藏主题都视为特征词的概率分布。

利用LDA模型获取科技政策文本隐藏主题的分布概率,其步骤如下。

输入:生成的主题个数 K;科技政策数据集 $\mathcal{D}$。

输出:科技政策数据集中隐藏的主题信息。

Step 1:对 $\mathcal{D}$ 去停用词,提取文本中的有意义字符串。

Step 2:利用Gibbs采样发现主题。对主题集合 $z=\{t_1,t_2,\cdots,t_n\}$ 的每个元素进行初始化,初始化值为1到 K 之间的任意数字,获得马尔科夫链的初始状态。

Step 3:循环主题集合 z,对主题单词进行分配,获得马尔科夫链的下一个状态。

Step 4:不断重复Step 3,直至马尔科夫链接近目标分布,保存 z 的值;利用单词的后验概率,间接地推算出每个主题出现的概率以及每个单词出现的概率。

Step 5:得到科技政策集的隐藏主题信息。

1.5.2 基于BM25模型的主题-文本相关性特征获取 BM25是一种基于概率检索模型的算法,常被用于计算检索词与检索内容的相似度。该算法首先对用户输入的检索词进行解析,得到解析后的语素集合 $q=\{q_1,q_2,\cdots,q_n\}$。然后对于检索结果文档 D,计算每个语素 q_i 与文档 D 的相似度得分。将 q_i 与文档 D 的相似度得分进行加权求和,最终获得检索词 q 与文档 D 的相似度。

利用BM25算法计算检索词与文本相关性特征,其步骤如下。

输入:检索词 q;科技政策数据集 $\mathcal{D}$。

输出:检索词与每篇科技政策的相关度得分。

Step 1:对 $\mathcal{D}$ 去停用词,提取文本中的有意义字符串。

Step 2:将检索词 q 分解成语素 $\{q_1,q_2,\cdots,q_n\}$。

Step 3:分别计算每个语素 q_i 与文档 D 各层级的相关性得分。

Step 4:将每个语素与文档 D 的相关性得分进行加权求和。

Step 5:计算得到最终的查询词 q 与文档 D 的相关性分数。

Step 6:重复Step 3~5,最终计算出检索词 q 与所有文档的相关度得分。

1.6 文本距离与扩散经验值获取

文本距离表示针对某一主题的文本差异性。对于某一主题下的政策文本来说,其文本距离往往与扩散概率存在一定的相关性,文本距离越近,其扩散概率往往也越大,也就越可能存在扩散关系。

将提取的文本特征输入Ranking SVM模型进行训练,训练好的模型可以用于文本排序。输入检索词后,模型会按顺序输出文本序列及排序得分。将输出的文本评分的差值作为文本距离 L,

$$L(A,B)=R_{\text{score}}(A)-R_{\text{score}}(B), \tag{9}$$

其中:$L(A,B)$ 表示文本 A 与文本 B 之间的文本距离;$R_{\text{score}}(A)$ 和 $R_{\text{score}}(B)$ 分别表示文档 A 和文档 B 的排序评分。为了判断文本距离与扩散概率之间的关系,利用Pearson算法对二者进行了相关性计算。通过实验验证了文本距离与扩散概率具有高度的相关性,文本距离越小,则越可能存在扩散关系。

2 科技政策扩散识别实验

2.1 科技政策收集

以《国家中长期科学和技术发展规划纲要(2006—2020)》发布后2006—2021年中央及地方政府制定的科技政策为研究对象,主要收集在中国政府网、科技部及省市政府网站、中国科技情报网公开的具有正式文号的规范政策文本,政策的类型主要选取法律、条例、纲要、规划、计划、办法、决定、意见、细则、通知等,不包括回函、批示、领导讲话以及名单、行业标准等类型的政策,共收集科技政策8 435篇,并构建了小型科技政策数据库对资源进行存储。采集的科技政策的区域分布数量如表1所示。

2.2 有意义字符串发现

点间互信息(pointwise mutual information,PMI)可以用来度量两个词之间彼此依赖的程度,能有效地判断组合词内部的紧密程度。随机选取100篇科技政策文本作为测试语料,使用jieba分词工具对文本进行初始分词,对其中的有意义字符串进行提取。

2.2.1 阈值判断实验 为了判断PMI-Entropy方法的阈值,设置不同的阈值对有意义字符串进行提取,实验结果如表2所示。可以看出,当阈值为0.1时准确率和召回率相对较高,因此将有意义字符串提取时PMI-Entropy方法的阈值设为0.1。

表 1　科技政策区域分布数量

Table 1　Number of regional distribution of science and technology policies

发布主体	采集数量	发布主体	采集数量	发布主体	采集数量
国家	361	湖北	261	广西	203
浙江	532	湖南	253	重庆	177
北京	506	贵州	239	河北	173
上海	436	青海	237	云南	164
安徽	398	黑龙江	235	山东	163
福建	354	吉林	228	辽宁	143
天津	345	江西	215	宁夏	104
江苏	332	甘肃	213	海南	94
广东	297	山西	208	新疆	75
陕西	283	内蒙古	206	西藏	24
河南	271	四川	204		

表 2　PMI-Entropy 方法的阈值判断实验结果

Table 2　Experimental results of threshold judgment of PMI-Entropy method

阈值	准确率	召回率	*F*1
0.3	0.716 5	0.693 3	0.704 7
0.2	0.688 9	0.733 3	0.710 4
0.1	0.710 5	0.771 4	0.739 7
0.05	0.608 9	0.785 1	0.685 9

2.2.2　对比实验　为了更好地验证所提方法的有效性,将单独使用 PMI-Entropy、单独使用规则以及规则+ PMI-Entropy 三种方法进行对比,实验结果如表 3 所示。可以看出,使用 PMI-Entropy 方法的召回率较高,但准确率较低,这是由于提取出许多不符合政策领域语义表达的非有意义字符串;单独使用规则提取的方法虽然准确率较高,但不能覆盖所有的有意义字符串;本文采用的 PMI-Entropy 与规则相结合的方法可以有效地弥补两种方法的不足,其准确率和召回率均相对较高。

表 3　不同方法的对比实验结果

Table 3　Comparative experimental results of different methods

方法	准确率	召回率	*F*1
PMI-Entropy	0.715 2	0.765 4	0.739 4
规则	0.811 3	0.704 3	0.754 0
规则+PMI-Entropy	0.889 2	0.746 8	0.811 8

2.3　政策扩散特征提取实验

实验随机选取了具备承继关系的 500 个政策扩散对作为正例,不具备承继关系的 500 个政策扩散对作为反例,对政策对进行随机编号。然后对这 1 000 个科技政策对进行预处理,并提取组织结构和语义结构,分别计算其组织结构相似度、语义结构相似度、关键词承继度与文本相似度。

为了更好地对不同的相似性结果进行比较,对组织结构相似度进行了归一化处理。四个特征的计算结果如表 4 所示。可以看出,对比组织结构特征相似度和语义结构相似度,发现后者的数值相对较高,这是由于在计算语义结构相似度时,根据结构词表补充了语义内容,使语义内容更加丰富。同理,由于融入了更多的语义,基于 Doc2vec 的文本相似度比关键词承继度的值更高一些。

表 4　四个特征的计算结果

Table 4　Calculation results of four features

文档序号	组织结构相似度	语义结构相似度	关键词承继度	基于 Doc2vec 的文本相似度
1	0.371 6	0.385 8	0.274 4	0.331 0
2	0.064 9	0.264 5	0.289 5	0.261 5
3	0.476 7	0.469 4	0.658 0	0.679 1
4	0.318 7	0.595 1	0.298 8	0.578 0
5	0.284 1	0.324 9	0.278 2	0.342 1
⋮	⋮	⋮	⋮	⋮
999	0.374 1	0.374 1	0.495 2	0.583 2
1 000	0.104 0	0.163 6	0.314 8	0.341 8

2.4　政策扩散识别模型训练与优化

以“知识产权”为检索词,在自建语料库中共检索出 427 篇政策,下述实验主要针对这些政策文本。在构建识别模型之前,需要对特征进行提取。由于前文已经涉及了组织结构相似度、语义结构相似度、关键词承继度和文本相似度特征提取的实验内容,在此重点介绍基于 LDA 的主题特征提取和基于 BM25 的主题-文本相关性特征提取。

提取 LDA 主题特征时,需要对 LDA 的主题数量进行确定。在实验中设定主题数分别为 8、10、12、14、16,迭代次数分别为 100、200、…、1 000 进行训练。在迭代次数相同而主题数不同时,平均子主题相似度情况如图 7 所示。

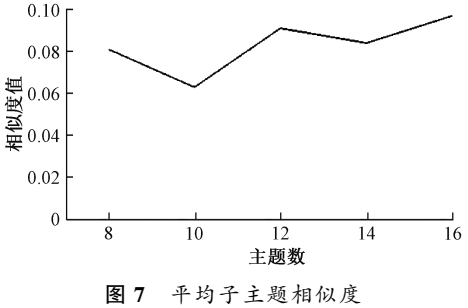

图 7　平均子主题相似度

Figure 7　Average subtopic similarity

从图7可以看出,主题数为10时相似度最低。关于知识产权的前5个高概率子主题列举如下。

topic 0:维权、知识产权行政、中介机构、重点产业、技术创新、集群、利益、社会公众、经济区、数据库。

topic 1:旅游、种质资源、战略纲要、国际旅游岛、专利技术、机构、奖励、重点优势产业、海洋、人民政府。

topic 2:产业、战略优势、高等院校、强省、发展、科研机构、关键技术、布图、经济社会。

topic 3:发展、企业、制度、产业、知识产权、创新、政府、文化、知识产权、政策。

topic 4:专利、信息、信息利用、企业、专利产业化、产业工作、商标注册量、专利转让、试点工作、申请。

接着对427篇政策文本的BM25特征进行提取,实验结果如表5所示。可以看出,由于标题字段文本较短,经过预处理后包含检索词的文献个数较少,经计算后其得分值为正数;而正文字段由于其文本较长,经过预处理后包含检索词的文献个数较多,其得分值为负数。同时发现标题字段、摘要字段以及正文字段与查询词的BM25得分具有一致性。

表5 对BM25特征进行提取的实验结果

Table 5 Experimental results of extracting BM25 features

文档序号	标题字段得分	正文字段得分
1	0.943 1	-9.442 8
2	2.539 0	-8.651 1
3	2.326 3	-9.648 7
4	1.006 7	-10.058 3
⋮	⋮	⋮
426	0.989 2	-10.406 5
427	3.418 2	-8.310 4

输入特征值,利用Ranking SVM进行训练。为了满足算法需求,结合上传到文档储存与检索引擎(Solr)中科技文献的排序,构造正负样本。算法的平均子主题召回率如图8所示。可以看出,当主题数为15时,平均子主题召回率最高。

对政策文本计算文本距离和扩散概率相关性,结果表明,Pearson系数为0.432 513;显著值为0.003 413,文本距离与扩散概率具有一定的相关性。之后对文本距离和扩散经验值进行计算,L_1为同一子主题的扩散经验值,L_2为不同子主题的扩散经验值,如果同主题$L_{i,j} < L_1$,不同主题$L_{i,j} < L_2$,则判断为具有扩散关系,加入扩散关系集中,计算使扩散关系成立的最大文本距离。最终测定当文本距离

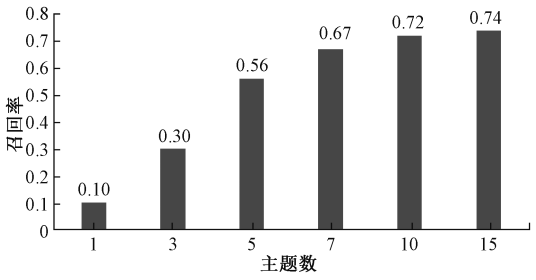

图8 平均子主题召回率

Figure 8 Average subtopic recall rate

$L < 0.035$时,基本可以判断具有扩散关系。

2.5 科技政策扩散识别展示

通过加入扩散经验值的识别模型,可以提取出科技政策扩散集合。以检索词"知识产权"为例,科技政策扩散集合展示界面如图9所示。

图9 科技政策扩散集合展示界面

Figure 9 Exhibition interface of science and technology policy diffusion set

3 小结

本文构建了科技政策扩散识别框架,根据政策间的文本距离与扩散概率的关系,寻找使扩散关系成立的最大文本距离。通过采用PMI-Entropy与规则相结合的方法,对政策领域中有意义字符串进行提取。通过对科技政策文本本身组织结构的挖掘,实现了对科技政策文本形式语义的解析,并在此基础上建立了科技政策领域词表来标引文本语义结构,实现了对文本语义的扩展。对科技政策扩散识别技术进行了研究,通过引入识别模型,实现多个特征的一体化处理,将分类与排序相结合,使最终提取的政策扩散集合能尽可能地覆盖检索主题内容,从而帮助用户迅速获取相关科技政策扩散情况,提高

分析效率。由于有意义字符串提取流程较长,且在提取过程中会过滤一些领域词汇,如何简化有意义字符串提取流程,提高准确率和召回率,是今后的研究工作之一。

参考文献:

[1] WALKER J L. The diffusion of innovations among the American states[J]. American political science review, 1969, 63(3): 880-899.

[2] 裴雷, 孙建军, 周兆韬. 政策文本计算: 一种新的政策文本解读方式[J]. 图书与情报, 2016(6): 47-55.
PEI L, SUN J J, ZHOU Z T. Policy text computing: a new methodology of policy interpretation[J]. Library & information, 2016(6): 47-55.

[3] 张剑, 黄萃, 叶选挺, 等. 中国公共政策扩散的文献量化研究: 以科技成果转化政策为例[J]. 中国软科学, 2016(2): 145-155.
ZHANG J, HUANG C, YE X T, et al. Study on China's public policy diffusion based on the quantitative analysis of policy documents: a case study on policies promoting commercialization of scientific and technological achievements[J]. China soft science, 2016(2): 145-155.

[4] 裴雷, 张奇萍, 李向举, 等. 中国信息化政策扩散中的政策主题跟踪研究[J]. 图书与情报, 2016(6): 63-71.
PEI L, ZHANG Q P, LI X J, et al. A statistical analysis of topic tracking in informatization policy diffusion in China[J]. Library & information, 2016(6): 63-71.

[5] 武学振. 中国省级政府信息政策创新扩散研究[D]. 南京: 南京大学, 2016.
WU X Z. A study on diffusion of innovation of information policy in provincial government in China[D]. Nanjing: Nanjing University, 2016.

[6] GRIMMER J, STEWART B M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts[J]. Political analysis, 2013, 21(3): 267-297.

[7] 李江, 刘源浩, 黄萃, 等. 用文献计量研究重塑政策文本数据分析: 政策文献计量的起源、迁移与方法创新[J]. 公共管理学报, 2015, 12(2): 138-144, 159.
LI J, LIU Y H, HUANG C, et al. Remolding the policy text data through documents quantitative research: the formation, transformation and method innovation of policy documents quantitative research[J]. Journal of public management, 2015, 12(2): 138-144, 159.

[8] NOWLIN M C. Modeling issue definitions using quantitative text analysis[J]. Policy studies journal, 2016, 44(3): 309-331.

[9] 王小杰. 政策扩散视角下中国铁路技术规章管理的文献量化与博弈研究[D]. 北京: 北京交通大学, 2018.
WANG X J. Document quantification and game research on technical regulation management of China railway under the perspective of policy dispersal[D]. Beijing: Beijing Jiaotong University, 2018.

[10] 李庆. 科技创新政策的转移、转移网络和竞争力研究: 以国家自主创新示范区为例[D]. 合肥: 中国科学技术大学, 2017.
LI Q. Research on science and technology innovation policy's transfer, network and competitiveness: a case study of national independent innovation demonstration zone[D]. Hefei: University of Science and Technology of China, 2017.